MANUEL D'UTILISATION DE L'OUTIL D'ETIQUETAGE SEMI-AUTOMATIQUE DES LISTES DE MOTS PFC

Cyril Auran et Jean-Michel Tarrier
ERSS UMR 5610 CNRS & Université de Toulouse-Le Mirail

0. Introduction

Ce manuel a pour but de présenter l'outil semi-automatique utilisable dans le cadre du traitement des fichiers de lecture de listes de mots. Le lecteur se réfèrera utilement à Tarrier & Auran (2004, ce volume), pour une description détaillée du format de rendu afférent.

L'outil dont il va être question dans ce manuel constitue une solution possible aux problèmes posés par cette tâche spécifique de PFC qu'est l'annotation alignée des listes de mots. Dans cette optique, deux phases sont à distinguer :

- Dans un premier temps, il est indispensable d'identifier les portions du signal sonore qui correspondent aux couples *nombre-mot* de la liste, et uniquement à ces éléments. Nous qualifierons cette phase « phase de segmentation ».
- Une fois la première phase terminée, le traitement se poursuit par la transcription orthographique des couples *nombre-mot*. Nous qualifierons cette phase « phase d'étiquetage ».

L'outil pour le traitement semi-automatique des listes de mots du projet PFC peut être téléchargé sous la forme d'un paquetage depuis l'adresse du site PFC :

http://infolang.u-paris10.fr/pfc/

Ou bien depuis le site suivant, rubrique « Ressources » du menu « Recherche » :

http://www.lpl.univ-aix.fr/~auran/

Ce paquetage comprend un script Perl d'installation, un fichier texte PFC.labels et un script Praat, « PFC_Mots.praat », qui, une fois installé et configuré, accomplit chacune des phases mentionnées plus haut :

- Génération automatique d'un TextGrid vide dont les portions correspondent aux segments inter-pauses (SIPs) détectés ; l'utilisateur vérifie manuellement les frontières ainsi créées et marque :
  - d'un « X » tout passage à supprimer entre des couples *nombre-mot*
  - d'un « M » tout passage à supprimer à l'intérieur d'un couple *nombre-mot*
- Remplissage des portions vides du TextGrid généré avec le texte correspondant au numéro et au mot lu par le locuteur selon les normes du format de rendu.

Nous allons à présent détailler de manière précise les différentes étapes nécessaires au traitement des fichiers de lecture de listes de mots dans le cadre de PFC.

1. Installation du script PFC_Mots.praat

1.1. Installation automatique

L'utilisateur pourra faire en sorte que le script soit installé dans Praat et puisse être appelé lors de l'ouverture d'un fichier son correspondant à un enregistrement de lecture de liste de

mots. C'est dans ce cas le script Perl « configure_PFC_Mots.pl », second script du paquetage, qui sera utilisé.

- **Phase (préparatoire) 1 : installation de l'interpréteur Perl**

Pour pouvoir faire fonctionner le script « configure_PFC_Mots.pl » sous Windows[1], il est indispensable que l'utilisateur installe l'interpréteur Perl que l'on pourra trouver à l'adresse suivante (cliquer sur « Download » en haut à gauche de l'écran et se laisser guider) :

http://www.activestate.com/Products/ActivePerl/

L'utilisateur récupèrera un programme qu'il faudra ensuite exécuter sur la machine concernée afin d'installer automatiquement l'interpréteur Perl.

- **Phase 2 : décompression du paquetage**

Une fois le paquetage correspondant à la plateforme de l'utilisateur récupéré, il convient de le décompresser dans le répertoire des outils PFC. L'utilisateur se réfèrera à la documentation relative à son système d'exploitation concernant cette opération[2].

Il est à noter que sous Windows 95/98/NT/2000, l'installation d'un logiciel spécifique (du type Winzip) est indispensable pour la décompression du paquetage .zip. Windows XP est en revanche capable de décompresser le paquetage de manière native.

- **Phase 3 : installation du script praat**

Sous Windows :

- S'assurer que Praat n'est pas ouvert ;
- Se placer dans le répertoire contenant le script PFC_Mots.praat ;
- Double-cliquer sur configure_PFC_Mots.pl[3] ;
- Suivre les instructions qui s'affichent à l'écran.

Sous Unix/linux/Mac os X :

- S'assurer que Praat n'est pas ouvert ;
- Dans un terminal, se placer dans le répertoire contenant le script PFC_Mots.praat à l'aide de la commande « cd » ;
- Entrer la commande suivante : « perl configure_PFC_Mots.pl » ;
- Suivre les instructions qui s'affichent à l'écran.

Le script PFC_Mots.praat devrait avoir été configuré pour être utilisable dans Praat lors de la sélection d'un enregistrement de lecture de liste de mots. L'utilisateur pourra vérifier l'installation en ouvrant un fichier son quelconque au moyen de la commande « Read from file… » et en observant la liste des boutons de commande disponibles. Le panneau « Objects » de Praat devrait alors être similaire à celui représenté en figure 1 ci-dessous.

[1] Les machines fonctionnant sous Unix/linux/Mac os X contiennent en général un interpréteur Perl : aucune installation n'est alors nécessaire.
[2] Les utilisateurs d'Unix/linux/os X utiliseront par exemple la commande « tar –xvf » ; les utilisateurs Mac (os X et versions antérieures) pourront utiliser le logiciel Stuffit Expander.
[3] Si le lancement ne se produit pas, ouvrir le fichier avec « perl.exe » qui se trouve dans le sous-dossier « bin » à l'intérieur du dossier « Perl ».

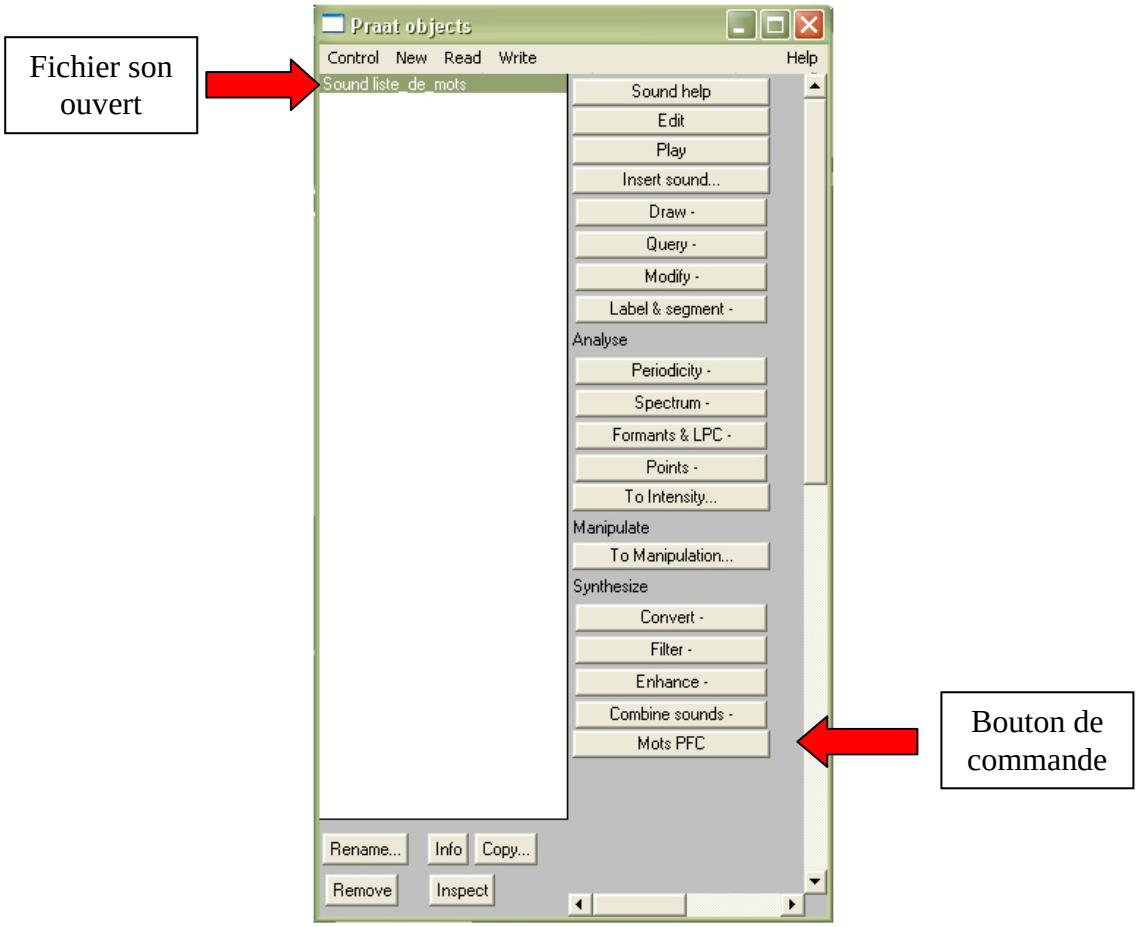

Figure 1 : Fenêtre « Objects » de Praat et bouton de commande « Mots_PFC »

Si un message d'erreur lors de l'installation automatique ou l'absence du bouton « Mots PFC » indique l'échec de l'installation automatique, l'utilisateur suivra alors la procédure manuelle d'installation présentée ci-après.

1.2. Installation manuelle (en cas d'échec de la phase 3)

1) Démarrer Praat
2) Sous « Control », cliquer sur « Open script… »
3) Sélectionner l'emplacement où se trouve le script PFC_Mots.praat
4) Sélectionner le script et cliquer sur « Ouvrir » ou appuyer sur Entrée
5) Sous « File… », cliquer sur « Add to dynamic menu… »
6) Ligne « Command », taper « Mots PFC »
7) Cliquer sur le bouton [OK]
8) Sous « File… », cliquer sur « Close »

Praat est à présent configuré pour pouvoir lancer le script PFC_Mots.praat lorsqu'un fichier son est sélectionné (cf. figure 1 ci-dessus).

2. Utilisation du script PFC_Mots.praat

2.1. Lancement

L'utilisateur pourra lancer le script à l'aide du bouton « Mots PFC » qui apparaît dans la fenêtre « Objects » de Praat lorsqu'un son est sélectionné. On se réfèrera utilement à

Delais-Roussarie *et al.*, 2002 pour des instructions relatives à l'ouverture d'un fichier son dans Praat.

2.2. Paramètres

Lors de l'exécution du script, la fenêtre ci-dessous apparaît à l'écran (figure 2) :

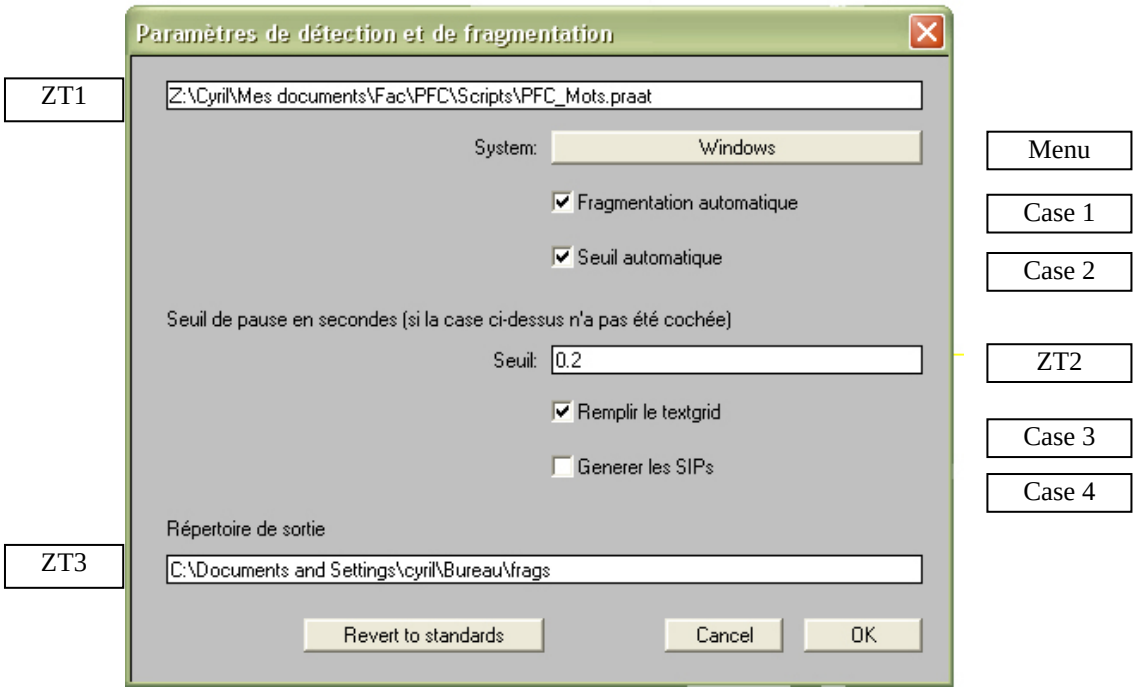

Figure 2 : Fenêtre de paramètres du script « PFC_Mots.praat »

Nous allons détailler les différentes zones de cette fenêtre en partant du haut :

- Zone de texte supérieure (ZT1) : Emplacement et nom du script ; ne pas modifier.
- Menu déroulant « System » (Menu) : Cliquer sur le bouton pour choisir le système approprié.
- Case « Fragmentation automatique » (Case 1) : cette option permet de choisir un algorithme automatique de segmentation (case cochée par défaut) ou d'opter pour une segmentation manuelle du fichier son (cf. § 2.3 ci-après).
- Case « Seuil automatique » (Case 2) : cette option n'est fonctionnelle que lorsque la case 1 est cochée ; elle permet d'effectuer un calcul automatique du seuil de détection des pauses (case cochée), ou bien d'opter pour un durée définie (case non cochée, voir ZT2).
- Zone de texte « Seuil » (ZT2) : cette zone permet de fixer le seuil de détection de pause (en secondes) lorsque la case 2 n'est pas cochée.
- Case « Remplir le TextGrid » (Case 3) : cette option permet de remplir les intervalles vides du TextGrid créé avec les couples « nombre-mot » de la liste (case cochée).
- Case « Generer les SIPs » (Case 4) : cette option permet de créer un fichier son pour chaque intervalle (ou « Segment Inter-Pause ») contenant un couple « nombre-mot » ; les fichiers son ainsi générés seront sauvegardés dans le répertoire de sortie

(cf. ZT3) et nommés d'après le modèle *nom-du-fichier-lecture-de-mot_seg01.wav* (pour le premier segment non marqué)

- Zone de texte « Répertoire de sortie » (ZT3) : Répertoire où les fichiers générés (TextGrid, fichier son épuré et éventuels SIPs) seront sauvegardés.

2.3. Fonctionnement

Les deux premiers aspects du fonctionnement du script PFC_Mots.praat concernent la phase de segmentation : la fragmentation et le nettoyage du fichier son y sont effectuées de manière semi-automatique. La dernière phase, « remplissage du TextGrid » correspond quant à elle à la phase d'étiquetage et peut être accomplie de manière totalement automatique.

- **Fragmentation automatique vs. fragmentation manuelle**

Etant donnée la nature « brute » de l'algorithme utilisé, il est indispensable de vérifier et (certainement) de corriger les TextGrids générés par le script lorsque l'option automatique est choisie. Cette correction intervient en cours de traitement, lorsque le script s'interrompt et rend temporairement la main à l'utilisateur.

On pourra se référer au manuel d'utilisation de Praat ou à l'introduction présente dans le numéro 1 du Bulletin PFC (Delais-Roussarie *et al.*, 2002) pour :

- supprimer (clic de sélection puis Alt+Back Space) ,
- plus rarement, déplacer (cliquez-glissez) ou bien,
- plus rarement encore, ajouter (clic puis Entrée)

les bornes des SIPs détectés automatiquement.

Si la case 1 n'est pas cochée, l'utilisateur devra ajouter manuellement les frontières délimitant les couples « nombre-mot » les uns des autres. Pour ce faire, il suivra la méthode suivante :

- lancer l'écoute à l'aide de la touche de tabulation du clavier ;
- ajouter une frontière par pression sur la touche [Entrée].

La fragmentation peut ainsi se dérouler quasiment en temps réel, les détails de positionnement pouvant ensuite être réglés par « cliquer-déplacer » à l'aide de la souris.

- **Nettoyage du fichier son**

Conformément aux recommandations données dans Durand *et al.* 2002 et reprises dans Eychenne *et al.* 2004, le fichier son final ne doit plus comporter « les commentaires qui précèdent et qui suivent la liste ou le texte à proprement parler » ; de plus, comme indiqué dans Tarrier & Auran 2004, les commentaires et remarques insérés dans la lecture de la liste doivent être supprimés eux-aussi.

Nous tenons à attirer tout particulièrement l'attention du lecteur sur l'importance des modifications apportées au fichier son lors de la phase de nettoyage. Dans l'optique de la conservation des données originales, nous recommandons avec Tarrier & Auran 2004 de déléguer cette tâche à un annotateur averti par point d'enquête qui pourra ainsi centraliser et systématiser le traitement des fichiers. Si toutefois les responsables d'un point d'enquête préfèrent déléguer cette phase du traitement aux annotateurs (utilisateurs des outils décrits

ici), il sera important de suivre scrupuleusement les indications données dans Tarrier & Auran 2004 et les instructions suivantes.

Lors de la phase de vérification de la fragmentation, l'utilisateur marquera dans la tire « SIPs » les intervalles à supprimer par les lettres « X » ou « M » (en majuscule) :

- On marquera d'un « X » toute portion à supprimer comprise entre deux couples *nombre-mot*.
- On marquera d'un « M » (comme « milieu ») toute portion à supprimer comprise à l'intérieur d'un couple *nombre-mot*.

Les portions de son correspondant à ces intervalles seront alors automatiquement supprimées du fichier son chargé dans Praat (le fichier d'origine ne sera pas modifié si le répertoire de sortie est différent du répertoire d'entrée).

A titre d'exemple, analysons le cas suivant emprunté à Tarrier & Auran 2004 :

> " 10 euh non c'est 11 ah non je ne me suis pas trompé excusez-moi mais je crois que je vais un peu vite je vais essayer de faire attention fêtard "

La totalité de la portion entre « 10 » et « fêtard » (c'est-à-dire « euh non c'est 11 ah non je ne me suis pas trompé excusez-moi mais je crois que je vais un peu vite je vais essayer de faire attention ») sera marquée d'un « M » car elle est interne au couple « 10 fêtard ».

En revanche si nous examinons :

> " 1 roc euh je ne sais pas si je lis bien ce qui est écrit mais reprenez-moi si ce n'est pas ce que vous voulez 2 rat",

tout ce qui est énoncé entre les deux lectures de mots (i.e. entre " 1 roc " et " 2 rat "), à savoir : " euh je ne sais pas si je lis bien ce qui est écrit mais reprenez-moi si ce n'est pas ce que vous voulez ", sera marqué d'un « X » (et ce, bien entendu, dans l'éventualité où une telle section n'aurait pas été supprimée dans le fichier sonore !).

Une fois les étapes de fragmentation et de nettoyage terminées, l'utilisateur déclenchera la suite du processus en cliquant sur le bouton « Continue » de la fenêtre « Pause » représentée dans la figure 3 ci-dessous.

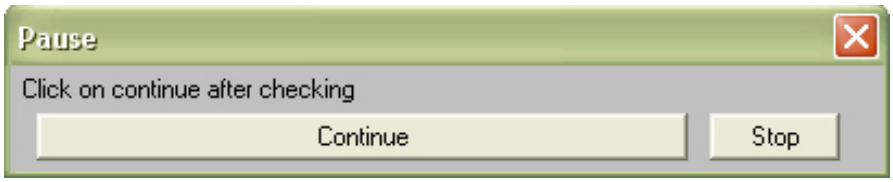

Figure 3 : Fenêtre « Pause »

- **Remplissage du TextGrid**

L'étape suivante du traitement consiste à remplir les intervalles vides du TextGrid (après vérification et correction manuelle) à l'aide des couples « nombre-mot » de la liste. Lorsque la case 3 est cochée (option par défaut), cette étape se déroule de manière totalement automatique et tout intervalle vide au-delà du 94ème est laissé vide.

- **Fin de traitement**

L'exécution du script se termine par la sélection simultanée du fichier son éventuellement épuré et du TextGrid correspondant ; l'utilisateur peut alors vérifier le traitement à l'aide de

la commande « Edit » avant de sauvegarder les données (fichier son puis TextGrid) dans le répertoire de son choix[4]. La sauvegarde elle-même sera effectuée par sélection exclusive de l'objet puis à l'aide des commandes du menu « Write ».

3. Evaluation du traitement semi-automatique

Pour apprécier toute l'utilité du script « PFC_Mots.praat », un test comparatif à été pratiqué entre, d'une part, la segmentation et l'étiquetage d'un fichier de manière entièrement manuelle et, d'autre part, ces mêmes opérations pour le même fichier mais cette fois-ci en recourant au traitement semi-automatique (avec les options « Fragmentation automatique », « seuil automatique » et « Remplir le TextGrid »). Pour information, ce test a été pratiqué sur un ordinateur fonctionnant sous Win2K et équipé d'un processeur « Duron » de 1 GHz et de 256Mo de mémoire RAM.

Procédure semi-automatique : moins de 17mn environ
- lancement du script : 1mn
- contrôle de la segmentation, correction et ajustage (insertion de « M » et « X », insertion ou effacement de balises) : 15 mn environ
- construction du fichier TextGrid final et étiquetage : quelques secondes

Procédure entièrement manuelle : 32mn environ
- segmentation et insertion de balises : 7mn
- Étiquetage (par copier/coller): 25mn

Ce petit test montre que l'utilisation du script permet une économie de temps et de manipulations plus qu'appréciable, et tout particulièrement pour l'opération d'étiquetage puisque là où la procédure automatique n'opère qu'en quelques secondes, il faut en revanche 25 minutes pour étiqueter manuellement par copier/coller les intervalles du fichiers.

Toutefois, pour ce qui concerne l'opération de segmentation il est possible de remarquer que, dans le test, la procédure manuelle s'avère plus rapide pour ces fichiers de liste de mots. En effet leur brièveté, de même que la répétitivité du train d'onde ainsi que sa grande lisibilité et « prédictibilité » rendent le découpage quasiment « automatique » dès lors que l'on y est familier. Chacun donc appréciera ici en fonction de ses performances. De fait, il pourrait être très intéressant de combiner à la fois segmentation entièrement manuelle et étiquetage automatique, ce que permet là encore le script PFC_Mots.praat.

Références

TARRIER J.-M., AURAN C. (2004). " Fichiers mots : constitution, alignement et transcription " (ce volume).

DELAIS – ROUSSARIE E., DURAND J., LYCHE C., MEQQORI A., TARRIER J.-M. (2002). " Transcription des données : outil et conventions " in Durand J., Laks B., Lyche C. (éds.) *Bulletin PFC n°1, Protocole, conventions et directions d'analyse*, pp 21-34. ERSS UMR 5610, CNRS & Université de Toulouse – Le Mirail.

DURAND J., LAKS B., LYCHE C. (2002). " Format des rendus 2002 et 2003 " in Durand J., Laks B., Lyche C. (éds.) *Bulletin PFC n°1, Protocole, conventions et directions d'analyse*, pp 71-74. ERSS UMR 5610, CNRS & Université de Toulouse – Le Mirail.

[4] Le fichier son et le TextGrid ont à ce stade déjà été sauvegardés automatiquement dans le répertoire de sortie défini en ZT3 (cf. figure 2).

DURAND J., LYCHE C., LAKS B., (2002). “ Protocole d’enquête ” in Durand J., Laks B., Lyche C. (éds.) *Bulletin PFC n°1, Protocole, conventions et directions d’analyse*, pp 7-19. ERSS UMR 5610, CNRS & Université de Toulouse – Le Mirail.

EYCHENNE J., HAMBYE P., MALLET G., (2004). “Format des rendus” (ce volume).